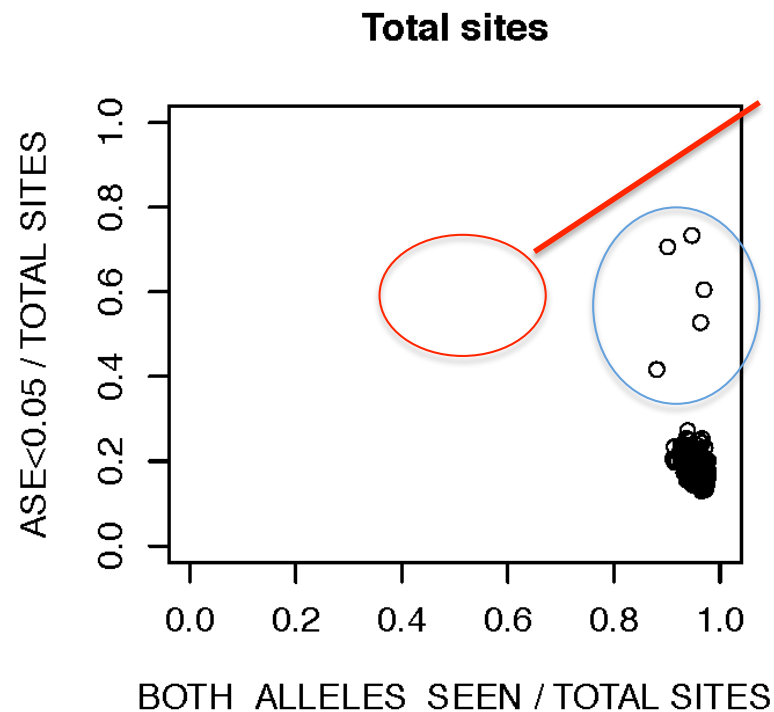


# Geuvadis RNAseq data Quality control analyses

Tuuli Lappalainen  
University of Geneva  
August 9th, 2012

# RNA – genotype concordance

- Using the ASE pipeline: RNAseq reads over heterozygous sites of each individual
  - both alleles should usually be observed in RNAseq data
  - the proportion of sites per individual with significant ASE should be relatively constant

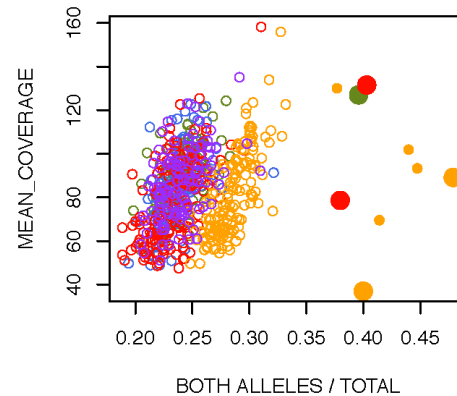
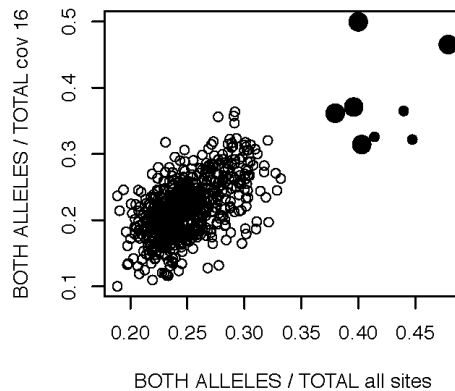
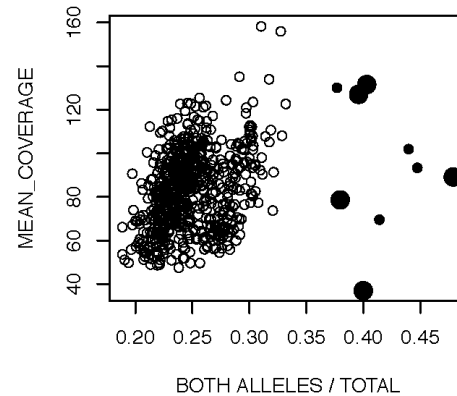
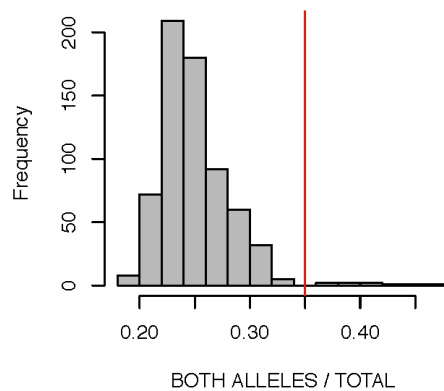


sample swaps  
would be here

why do these 5 samples have increased allelic  
imbalance?

- problem in the genotypes or in the RNA data?  
- could it be sample cross-contamination? A bit of  
RNA from a homozygote individual would bias the  
allelic ratios...

# Detecting cross-contamination



- Cross-contamination in RNAseq -> increased heterozygosity
- Run the ASE pipeline on all sites that are polymorphic in the studied population (chr1 sites > 5% MAF in EUR or YRI) – the same set of variants across all RNAseq samples
- The 5 outliers show increased heterozygosity
- 4 other samples with increased heterozygosity in this analysis
  - all YRI
  - could be a population genetic phenomenon – maybe these samples are for some reason more heterozygous for this set of SNPs

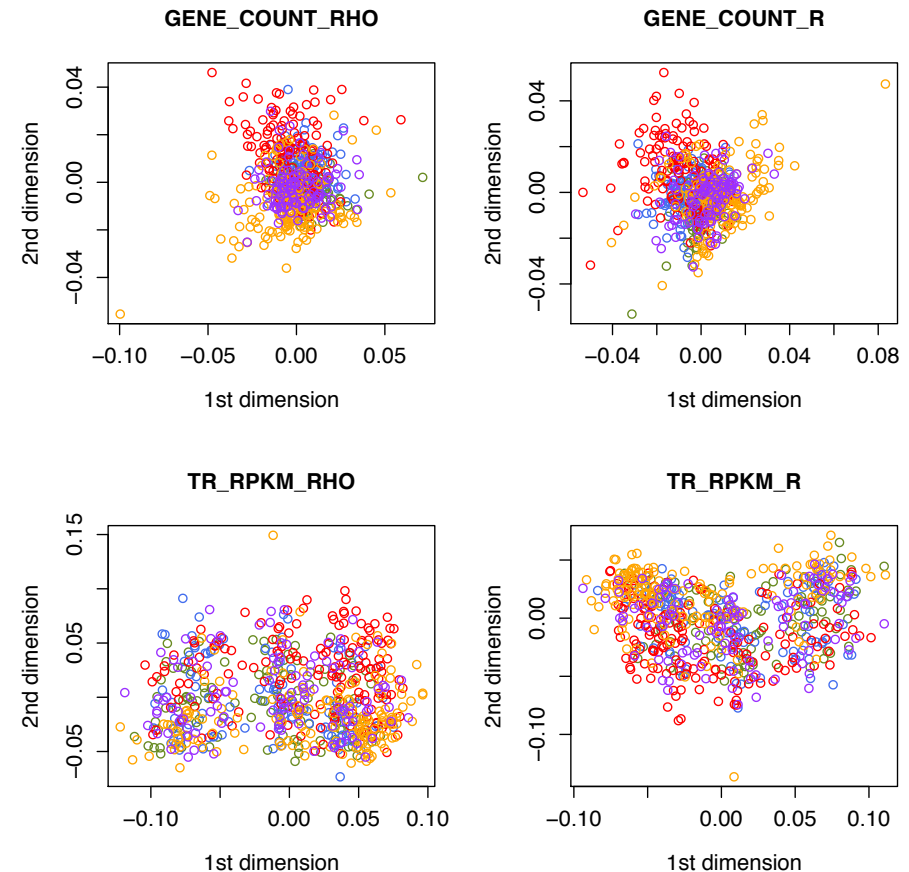
Big dots = the 5 outliers from the previous plot

# RNA – genotype matching: results

- We don't have a single sample swap in the 667 samples!
- The 5 outlier samples that have signs of cross-contamination – **exclude from all analysis**
  - NA19225.6.M\_120119\_5
  - NA12399.7.M\_120219\_1
  - NA07000.1.M\_120209\_2
  - NA18861.4.M\_120208\_5 (this sample fails in every analysis)
  - HG00237.4.M\_120208\_1
- The 4 other samples with some increased heterozygosity – include in analysis, but be a bit cautious
  - NA19095.5.M\_120131\_5
  - NA19130.1.M\_120209\_1
  - NA19144.4.M\_120208\_2
  - NA19235.1.M\_111124\_6

# Expression level QC analysis

- How to measure sample similarity?
  - Spearman rho
  - Optimal Power Space Transformation to scale quantifications and remove outliers + Pearson's r
    - R package ops written by Micha and Paolo
- Gene and exon read counts normalized by the total number of mapped reads
- Transcript read counts normalized by the total number of mapped reads, transcript RPKMs
- Todo: junction and intron quantifications
- OPS + Pearson's correlation between all sample pairs
  - random 10 000 elements
  - quantification >0 in both samples
- D-statistic per sample : correlation of that sample with all the others
- Multidimensional scaling of the distance matrix

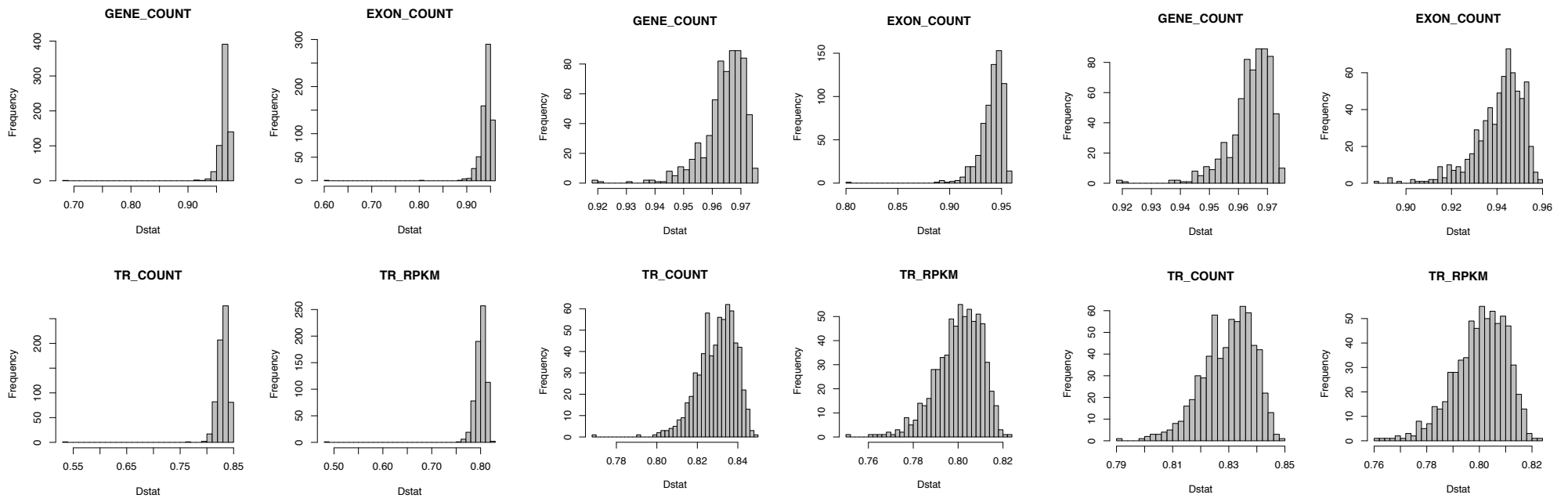


# D-statistic histograms

All 667

Excluded  
NA18861.4.M\_120208\_5

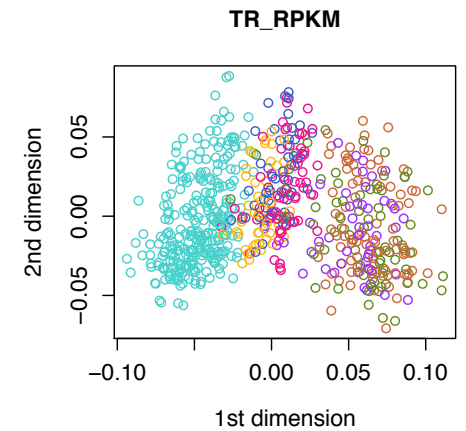
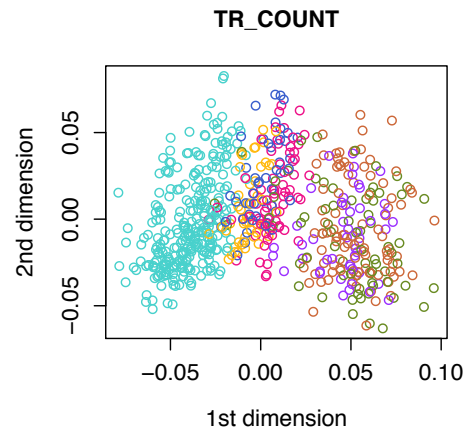
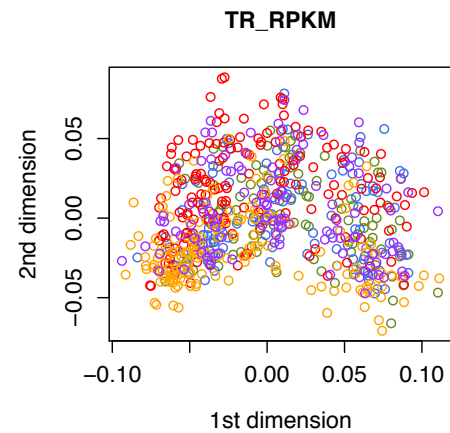
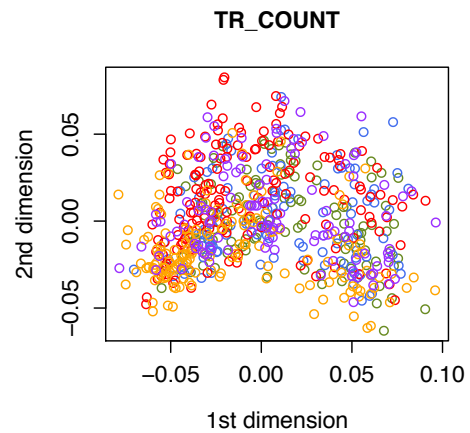
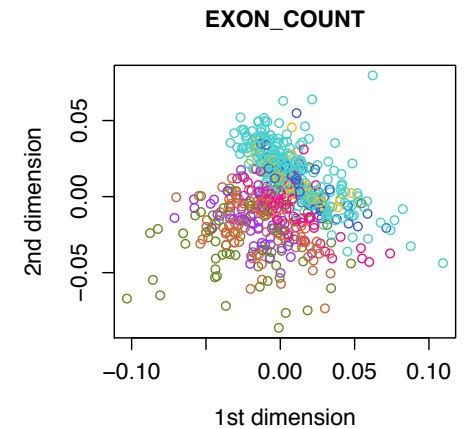
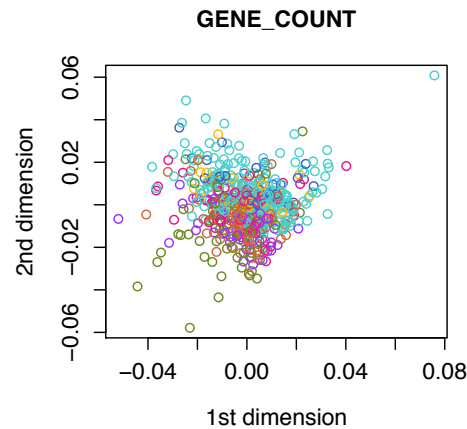
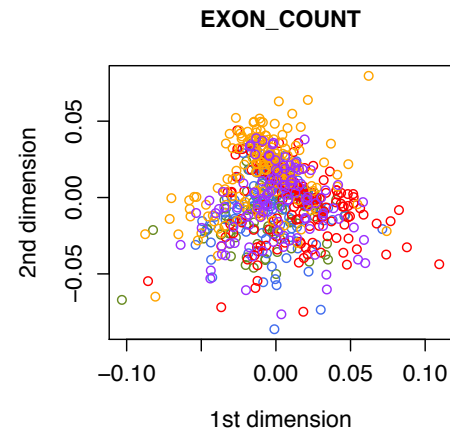
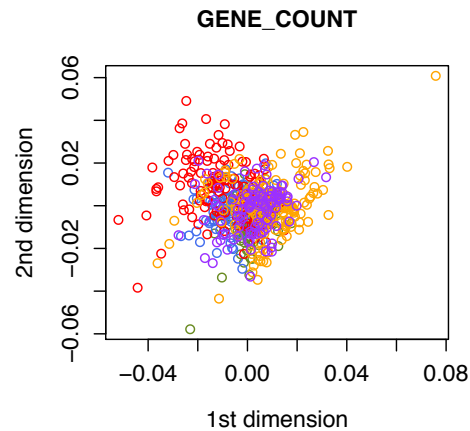
Excluded  
NA18861.4.M\_120208\_5  
NA19144.4.M\_120208\_2



# Sample clustering

Population colors

Lab colors



Excluded NA18861.4.M\_120208\_5, NA19144.4.M\_120208\_2

# Technical covariates of exon and transcript quantifications

- Covariates to test: UNIGE & Uppsala QC outputs, 237 stats in total (with some redundancy) – see next slide
- Analysis
  - Correlation between the covariate and expression levels of each exon/transcript -> p-value for each exon
  - Pi1 value to measure the enrichment of low p-values. High number = many exons' quantifications correlate with the covariate
  - Regression to remove the effect of covariates -> residuals used as new quantifications
  - Sequential correction of covariates based on the pi1 values and gut feeling...
- This type of correction can be quite subjective, but is informative for understanding technical sources of variation

SampleID	Qual_average_CT_count	GENE_COUNT	Fraction_A_for_cycles_25_C_4_9	Mean_copy_number	Gene_body_coverage_nr_5
Population	Qual_average_GA_count	EXON_COUNT	Fraction_A_Stddev_for_cycles_25_C_49	Mean_copy_number_stddev	Gene_body_coverage_perc_5
SeqLabNumber	Qual_average_GC_count	lane	Fraction_C_for_cycles_25_C_4_9	Total_reads	Gene_body_coverage_nr_6
RNAExtractionBatch	Qual_average_GG_count	datelane	Fraction_C_Stddev_for_cycles_25_C_49	Unmapped_reads	Gene_body_coverage_perc_6
RNAQuality	Qual_average_GT_count	Mean_Q_StdDev	Fraction_G_for_cycles_25_C_4_9	Percent_unmapped_reads	Gene_body_coverage_nr_7
RIN	Qual_average_TA_count	Fraction_N	Fraction_G_Stddev_for_cycles_25_C_49	Uniquely_mapped	Gene_body_coverage_perc_7
RNAConcentration_ng.ul	Qual_average_TC_count	Fraction_N_Stddev	Fraction_T_for_cycles_25_C_4_9	Percent_uniquely_mapped	Gene_body_coverage_nr_8
RNAQuantityLibraryPrep_ng	Qual_average_TG_count	Fraction_A	Fraction_T_Stddev_for_cycles_25_C_49	Diff_read1_and_read2	Gene_body_coverage_perc_8
LibraryPrepPlate	Qual_average_TT_count	Fraction_A_Stddev	Fraction_bases_with_Q2_for_cycles_50_C_74	Ratio_read1_read2	Gene_body_coverage_nr_9
LibraryPrepDate	Qual_bottom_10_perc_mean	Fraction_C	StdDev_Q_for_cycles_50_C_74	Diff_reads_map_to_for_and_re_v	Gene_body_coverage_perc_9
Operator	Qual_CA_mean_quality	Fraction_C_Stddev	Fraction_bases_with_Q10_for_cycles_50_C_74	Ratio_for_rev	Gene_body_coverage_nr_10
PrimerIndex	Qual_CC_mean_quality	Fraction_G	Fraction_bases_with_Q20_for_cycles_50_C_74	Non_splice_reads	Gene_body_coverage_perc_10
LibraryConcentrationMethod	Qual_CG_mean_quality	Fraction_G_Stddev	Fraction_bases_with_Q30_for_cycles_50_C_74	Percent_non_splice_reads	Gene_body_coverage_nr_11
LibraryConcentration_ng.ul	Qual_CT_mean_quality	Fraction_T	Fraction_N_for_cycles_50_C_7_4	Splice_reads	Gene_body_coverage_perc_11
BioanalyzerSize_bp	Qual_GA_mean_quality	Fraction_T_Stddev	Fraction_N_Stddev_for_cycles_50_C_74	Percent_splice_reads	GC_in_max_content
LibraryQuantitySequencing_p_M	Qual_GC_mean_quality	Mean_Q_for_cycles_0_C_24	Fraction_A_for_cycles_50_C_7_4	Percent_reads_mapped_in_proper_pairs	Percent_GC_in_max_content
ClusterDensityRaw	Qual_GG_mean_quality	StdDev_Q_for_cycles_0_C_24	Fraction_C_for_cycles_0_C_24	Clipping_profile_nr_1	Total_normalized_difference_of_nucleotide_content_towards_normal
ClusterDensityPass	Qual_GT_mean_quality	Fraction_bases_with_Q2_for_cycles_0_C_24	Fraction_C_Stddev_for_cycles_0_C_24	Clipping_profile_perc_1	percGC
SequencingDate	Qual_mean	Fraction_bases_with_Q10_for_cycles_0_C_24	Fraction_G_for_cycles_0_C_24	Clipping_profile_nr_2	Min_per_base_seq_quality
Machine	Qual_median	Fraction_bases_with_Q20_for_cycles_0_C_24	Fraction_G_Stddev_for_cycles_0_C_24	Clipping_profile_perc_2	Max_per_base_seq_quality
FlowCell	Qual_Ns_count	Fraction_bases_with_Q30_for_cycles_0_C_24	Fraction_T_for_cycles_0_C_24	Clipping_profile_nr_3	Max_GC_percentile
Lane	Qual_number_of_values	Fraction_N_for_cycles_0_C_24	Fraction_T_Stddev_for_cycles_0_C_24	Clipping_profile_perc_3	ebv
HCSversion	Qual_TA_mean_quality	Fraction_N_Stddev_for_cycles_0_C_24	Mean_Q_for_cycles_25_C_49	Clipping_profile_nr_4	ebv_prom
RTAversion	Qual_TC_mean_quality	Fraction_N_Stddev_for_cycles_0_C_24	StdDev_Q_for_cycles_25_C_49	Clipping_profile_perc_4	
GC_bottom_10_perc_mean	Qual_TG_mean_quality	Fraction_A_for_cycles_0_C_24	Fraction_bases_with_Q2_for_cycles_25_C_49	Clipping_profile_nr_5	
GC_mean	Qual_top_10_perc_mean	Fraction_A_Stddev_for_cycles_0_C_24	Fraction_bases_with_Q10_for_cycles_25_C_49	Clipping_profile_perc_5	
GC_stddev	Qual_TT_mean_quality	Fraction_C_for_cycles_0_C_24	Fraction_bases_with_Q20_for_cycles_25_C_49	Clipping_profile_nr_6	
GC_top_10_perc_mean	Qual_10_perc_count	Fraction_C_Stddev_for_cycles_0_C_24	Fraction_bases_with_Q30_for_cycles_25_C_49	Clipping_profile_perc_6	
GC_10_perc_count	INSERT_SIZE_MODE	Fraction_G_for_cycles_0_C_24	Fraction_N_for_cycles_25_C_4_9	Clipping_profile_nr_7	
Number_of_unique_15mers	Mapped	Fraction_G_Stddev_for_cycles_0_C_24	Mean_Q30_length	Clipping_profile_perc_7	
Perc_explained_by_top_1_15mers	Multiple_mapping	Fraction_T_for_cycles_0_C_24	Mean_Q30_length_Stddev	Clipping_profile_nr_8	
Perc_explained_by_top_10_15mers	Proper_pair	Fraction_T_Stddev_for_cycles_0_C_24	Mean_GC_content_of_sequences	Clipping_profile_perc_8	
Perc_explained_by_top_25_15mers	Q_EQUALS_0	Mean_Q_for_cycles_25_C_49	Mean_GC_content_of_sequences_Stddev	Duplicates	
Perc_explained_by_top_50_15mers	Q_GREATER_THAN_150	StdDev_Q_for_cycles_25_C_49		Percent_duplicates	
Perc_explained_by_top_100_15mers	Q_GREATER_THAN_150_and_proper_pair	Fraction_bases_with_Q2_for_cycles_25_C_49		Optical_duplicates	
	Read_one_mapped	Fraction_bases_with_Q10_for_cycles_25_C_49		Percent_optical_duplicates	
	X_Read_two_mapped	Fraction_bases_with_Q20_for_cycles_25_C_49		Percent_duplication	
	Read_unmapped	Fraction_bases_with_Q30_for_cycles_25_C_49		Estimated_library_size	
	Total_read	Fraction_N_for_cycles_25_C_4_9		Estimated_library_size_2	
	unique_and_Q_EQUALS_0	Fraction_N_Stddev_for_cycles_25_C_49		Nr_duplicates_less_10	
	Unique_mapping			Percent_duplicates_less_10	
	BAD_CODES			Gene_body_coverage_nr_1	
	BELOW_MAPPING_QUALITY			Gene_body_coverage_perc_1	
	Total_read_NM6_MAPQ150			Gene_body_coverage_nr_2	
	TOTAL_EXONIC			Gene_body_coverage_perc_2	
	TOTAL_EXONIC_OVER_TOTAL_READS			Gene_body_coverage_nr_3	
	Total_read_NM6			Gene_body_coverage_perc_3	
	UNIQUE_VALID_POSITIONS			Gene_body_coverage_nr_4	
				Gene_body_coverage_perc_4	

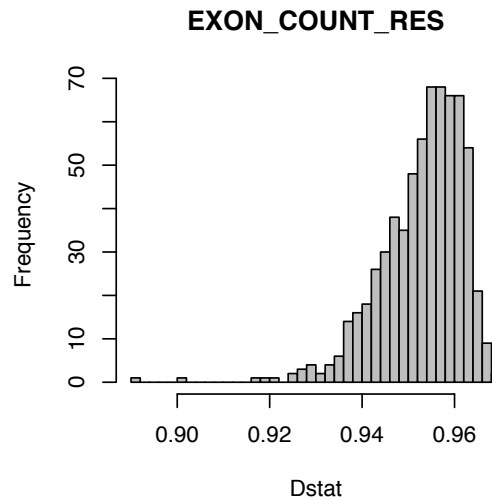
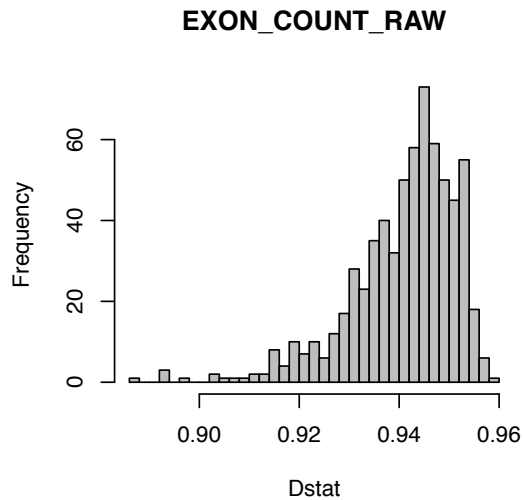
# Exon level top covariates

Raw data	Residuals after SeqLabNumber, GC_mean	Residuals after SeqLabNumber, GC_mean, RIN	Residuals after SeqLabNumber, GC_mean, RIN, TOTAL_EXONIC (reads)
Gene_body_coverage_perc_9 : 0.92	Population : 0.99	Gene_body_coverage_perc_9 : 0.87	Population : 0.98
Gene_body_coverage_perc_10 : 0.91	Gene_body_coverage_perc_9 : 0.91	Estimated_library_size : 0.87	RNAExtractionBatch : 0.97
Operator : 0.91	Estimated_library_size : 0.91	Clipping_profile_perc_2 : 0.86	Gene_body_coverage_perc_9 : 0.87
Gene_body_coverage_perc_8 : 0.90	Gene_body_coverage_perc_10 : 0.91	Estimated_library_size_2 : 0.85	Gene_body_coverage_perc_8 : 0.86
RIN : 0.89	RIN : 0.90	Gene_body_coverage_perc_8 : 0.85	Clipping_profile_perc_2 : 0.85
Percent_duplicates : 0.89	EXON_COUNT : 0.90	EXON_COUNT : 0.83	Estimated_library_size : 0.84
Percent_duplication : 0.89	Estimated_library_size_2 : 0.90	Percent_splice_reads : 0.80	EXON_COUNT : 0.82
Gene_body_coverage_perc_5 : 0.88	Percent_splice_reads : 0.90	Percent_non_splice_reads : 0.80	Estimated_library_size_2 : 0.82
Gene_body_coverage_perc_2 : 0.87	Percent_non_splice_reads : 0.90	Gene_body_coverage_perc_10 : 0.80	Percent_duplication : 0.81
EXON_COUNT : 0.87	Clipping_profile_perc_2 : 0.89	Perc_explained_by_top_100_15mers : 0.80	Percent_duplicates_less_10 : 0.81

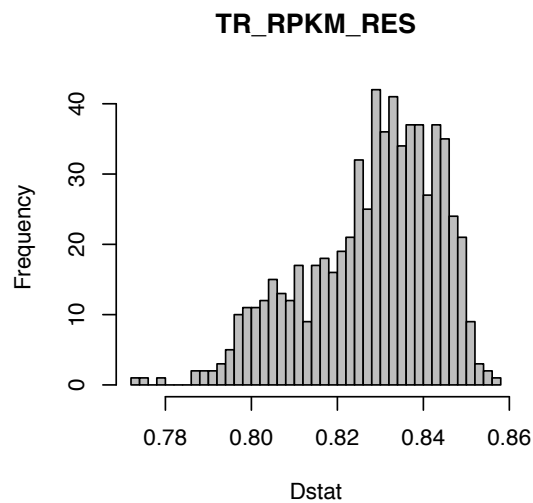
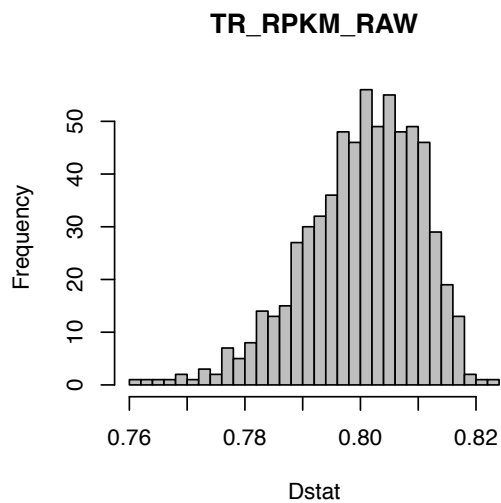
# Transcript RPKM top covariates

Raw data	Residuals after SeqLabNumber, GC_mean	Residuals after SeqLabNumber, GC_mean, RIN	Residuals after SeqLabNumber, GC_mean, RIN, TOTAL_EXONIC (reads)
SeqLabNumber : 0.92	Population : 0.76	Population : 0.69	Population : 0.67
LibraryPrepDate : 0.91	EXON_COUNT : 0.69	EXON_COUNT : 0.65	EXON_COUNT : 0.65
Operator : 0.86	Estimated_library_size : 0.69	GENE_COUNT : 0.63	GENE_COUNT : 0.64
LibraryPrepPlate : 0.85	Estimated_library_size_2 : 0.67	TOTAL_EXONIC_OVER_TOTAL_ READS : 0.63	TOTAL_EXONIC_OVER_TOTAL_ READS : 0.64
SequencingDate : 0.83	Percent_splice_reads : 0.67	Estimated_library_size : 0.62	Estimated_library_size : 0.63
Qual_average_CG_count : 0.82	Percent_non_splice_reads : 0.67	Number_of_unique_15mers : 0.62	Estimated_library_size_2 : 0.61
FlowCell : 0.81	Gene_body_coverage_perc_10 : 0.66	Estimated_library_size_2 : 0.62	Perc_explained_by_top_100_1 5mers : 0.60
Fraction_T_for_cycles_50_C_7 4 : 0.81	RIN : 0.65	UNIQUE_VALID_POSITIONS : 0.60	Perc_explained_by_top_50_15 mers : 0.60
Clipping_profile_perc_1 : 0.80	TOTAL_EXONIC_OVER_TOTAL_ READS : 0.65	Splice_reads : 0.59	Percent_splice_reads : 0.58
Qual_average_GC_count : 0.80	Number_of_unique_15mers : 0.64	Percent_splice_reads : 0.58	Percent_non_splice_reads : 0.58

# D-statistic pre and post normalization



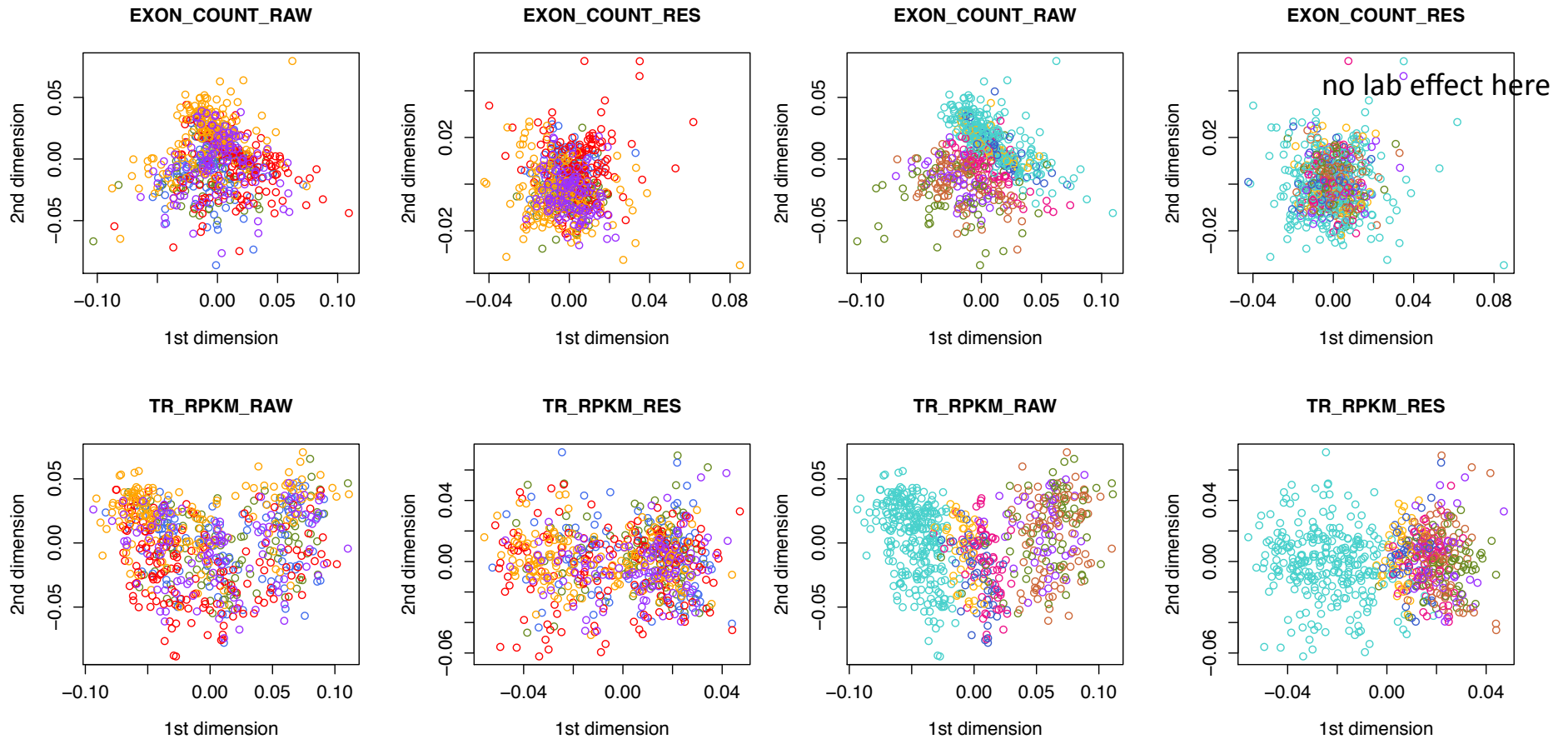
Exon correlation improves  
a lot, transcript not that  
much



# Sample clustering pre and post normalization

Population colors

Lab colors



big lab effect (unige vs others)  
still here...would need additional  
correction

# Covariates of Sequencing Lab

COVARIATE	-LOG10 p		
RNAQuantityLibraryPrep_ng	Inf	Qual_average_AG_count	160.97
Fraction_G_Stddev_for_cycles_50_C_74	Inf	Clipping_profile_nr_7	156.69
LibraryQuantitySequencing_pM	308.22	Fraction_bases_with_Q2_for_cycles_0_C_24	155.97
Fraction_C_Stddev_for_cycles_0_C_24	267.61	Optical_duplicates	154.37
<b>Clipping_profile_perc_8</b>	254.80	<b>INSERT_SIZE_MODE</b>	152.11
<b>GC_stddev</b>	233.54	Fraction_G_Stddev	151.33
Mean_GC_content_of_sequences_Stddev	231.56	Qual_average_TG_count	140.73
Clipping_profile_perc_7	224.76	Fraction_T_Stddev	132.98
<b>Qual_average_AC_count</b>	214.47	LibraryConcentration_ng.ul	132.62
Fraction_T_Stddev_for_cycles_0_C_24	208.78	BioanalyzerSize_bp	130.93
Qual_average_TC_count	205.95	Fraction_C_for_cycles_50_C_74	130.53
Fraction_C_Stddev	201.98	ClusterDensityPass	129.89
Fraction_T_Stddev_for_cycles_50_C_74	200.96	Fraction_bases_with_Q10_for_cycles_0_C_24	125.47
Qual_average_GA_count	196.05	ClusterDensityRaw	123.17
Fraction_A_Stddev_for_cycles_50_C_74	177.32	Fraction_bases_with_Q2_for_cycles_50_C_74	119.27
Qual_average_CT_count	174.62	Qual_average_GC_count	114.90
Qual_average_CA_count	171.53	Qual_average_CG_count	106.27
Qual_average_GT_count	163.94	Fraction_bases_with_Q10_for_cycles_50_C_74	99.95
<b>Percent_optical_duplicates</b>	162.72	Fraction_T_for_cycles_50_C_74	99.75